

DOCUMENT RESUME

ED 158 374

EA 010 772

AUTHOR Shann, Mary H.
TITLE Assessing Accountability for Basic Skill Development in the Adoption of a New Curriculum.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Mar 78
GRANT PES-74-00542-A01
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (Toronto, Ontario, March 27-31, 1978)

EDRS PRICE MF-\$0.83 HC-\$1.67 Plus Postage.
DESCRIPTORS *Basic Skills; Educational Accountability; Elementary Secondary Education; *Mathematics Instruction; Problem Solving; *Science Instruction; Testing Problems
IDENTIFIERS Stanford Achievement Tests; Unified Science Mathematics for Elementary Schools

ABSTRACT

The Unified Science and Mathematics for Elementary Schools (USMES) Program was designed to teach students complex problem-solving skills. This paper reports the effects of USMES on basic skills, explains the effects, comments on difficulties in data collection, and recommends data-gathering that should complement or replace standard basic skills test scores for curriculum decisions. Classes from 15 geographic areas were used to make up a national sample of classrooms using the USMES program. Control classes were chosen from the same communities as the USMES schools or from neighboring communities. The USMES classes performed at least as well as the control matches on the six selected subtests of the Stanford Achievement Test (SAT). These results indicate that USMES is not interfering with basic skills acquisition. Data-gathering problems arose because many principals and teachers objected to the time demands for SAT administration. In some areas, state and local testing regulations were severely limiting. Another problem was the question of whether scaled scores are comparable across test levels in a sequential battery. Since the program was designed to teach problem-solving skills, program effectiveness should be assessed with objective instruments to measure complex problem-solving.

(Author/JM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED158374

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

Assessing Accountability for Basic Skill Development
in the Adoption of a New Curriculum

Mary H. Shann

Boston University

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Mary H. Shann

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

Paper presented at the Annual Meeting of the American Educational
Research Association, Toronto, March, 1978.

This evaluation was prepared with the support of National Science
Foundation Grant No. PES 74-00542 A01. Any opinions, findings,
conclusions or recommendations expressed herein are those of
the author and do not necessarily reflect the views of the
National Science Foundation.

EA 010 772

Assessing Accountability for Basic Skill Development in the
Adoption of a New Curriculum

MARY H. SHANN, Boston University

The innovative curriculum for complex problem solving called Unified Science and Mathematics for Elementary Schools (USMES) commenced trial implementation during the "back-to-basics" movement. Concern for USMES students' basic skill development required that the USMES evaluation include an assessment of the effects of USMES on basic skills. The purposes of this paper are to report those effects, to suggest why they occurred, to comment on the difficulties encountered in the data collection and analysis for this assessment, and to recommend what kinds of data need to complement or even replace standardized basic skills test scores for curriculum decisions.

Assessing Accountability for Basic Skill Development in the Adoption of a New Curriculum

The economic and political climate which ushered the era of accountability in education in the 1970's forced educators to take especially cautious views of new programs which could be regarded as "nice but not necessary," particularly if the programs might detract from efforts to teach basic skills. It was in this climate that the innovative curriculum for complex problem solving called Unified Science and Mathematics for Elementary Schools (USMES) commenced trial implementation. The National Science Foundation which funded the development of the USMES program also sponsored the independent evaluation of USMES, and a significant portion of the efforts to evaluate USMES involved assessing the effects of USMES on students' basic skill development. The purposes of this paper are to report those effects, to suggest why they occurred, to comment on the difficulties encountered in the data collection and analysis for this assessment, and to recommend what kinds of data need to complement or even replace standardized basic skills test scores as inputs for decisions on adopting educational programs.

Defining Accountability--In Theory and in Practice

Educational accountability has been defined in a myriad of ways with varying scope and emphases. Some writers limit their notion of the concept to the task of establishing clear, appropriate goals and examining results (Locke, 1971). To others, it is the process of evaluating programs but in a cost-analysis framework (Bell, 1971).

Disseminating the information to the public--informing the taxpayer what results at what costs--is an additional element of accountability for most users of the term (Cunningham, 1971; Gooler, 1971).

Also germane to the concept is the issue of who is accountable, and this is perhaps the most sensitive facet of this controversial movement.

Gallup's (1971) survey of attitudes toward public education estimated that 54% of the American public would place the chief blame for children's failures in school on the children's home life; only 8% would level primary responsibility with the teachers. When confronted with the same issue, 51% of a sample of high school juniors and seniors held that students themselves were chiefly to blame for their own academic failures, while 11% placed the chief blame on their teachers. Despite these statistics, the perception widely held by educators and reinforced by most writers on educational accountability is that teachers have been held as the primary accounting party (Mehrens & Lehmann, 1973).

The most fundamental aspect of any of the definitions offered for the term accountability involves the initial question of goal setting--accountability for what. Mehrens and Lehman (1973, p. 623) commented further on the desirability and feasibility of pursuing different kinds of outcome objectives: (a) minimally acceptable levels of proficiency for all students; (b) maximum productivity from students as a total group; (c) equal performance thus enabling equal opportunities for students as they commence the post-schooling portions of their lives; or (d) maximum development of each individual student

according to his potential. Dyer (1970) addressed "accountability for what" in a different way, emphasizing the responsibilities of the professional staff of a school "for keeping themselves informed about the diverse needs of their pupils and for doing the best they can to meet those needs" (1970, p. 206).

The complexity of very legitimate philosophical and political concerns which should underlie the conduct of accountability programs do not seem to be reflected in practice, however. The "back-to-basics" bandwagon which has swept the country during the 1970's coupled with the relative ease of measuring basic skills objectives undoubtedly account for the tendency to focus accountability programs narrowly on lower level cognitive goals, while excluding attention to the difficult-to-measure higher mental processes and to elusive affective goals. Thus, many educational systems have found themselves on the defensive, treating several dimensions of the accountability concept as purely philosophical, academic, or ethical concerns. Instead, they pursue its operational definition as the process of justifying educational expenditures by presenting evidence from standardized tests that their students are developing proficiencies in the basic skills areas of reading and arithmetic, seen as the enablers for the development of competencies in life-roles.

The USMES curriculum did not escape these pressures. In response to concerns voiced by current and prospective USMES for remaining successfully accountable for their students' basic skill development, the USMES developers and, in turn, the USMES evaluators had to seek

evidence that USMES students performed at least as well as non-USMES students in the basic skills.

The Role of Basic Skills in USMES

The USMES project's Central Staff at the Education Development Center in Newton, Massachusetts has sought to work with classroom teachers and specialists on the development and implementation of approximately 30 interdisciplinary units as examples of the real problem solving style of learning. The units are designed to engage elementary school students in long-range, complex investigations of real and practical problems taken from their school or community environment. By acting on these problems, called "challenges," the student is supposed to gain competence in problem solving, and do so in a manner that gives him an experiential understanding (learning-by-doing) of the problem solving process, as well as the acquisition of basic skills and concepts.

Thus, the paramount goal of the USMES program is the enhancement of students' abilities in real problem solving. However, two important criteria for the selection of an USMES unit challenge have been that the problem has the potential for a substantial acquisition of facts and scientific concepts and also for mathematical structuring appropriate to the age level (Education Development Center, September, 1972, p. 5). Additionally, the USMES developers have suggested that the group communication required for the solution of USMES challenges would enhance language arts skills essential to improvement in other cognitive areas.

One aspect of the original conception of USMES was that of an integrated mathematics, science, and social science program. Inherent in this conception was the notion that these disciplines, particularly mathematics and science, could be taught not just as discrete skills but by an integrated problem solving approach. In their concern that they communicate the skills and content areas of these disciplines to their students at least as well as they had with their more traditional approaches, USMES teachers and principals, as well as prospective USMES users asked: "Will students' scores in these areas decrease if these subjects are no longer taught as separate content areas?"

Over the next few years, the emphasis of USMES had been modified gradually. USMES no longer claimed to offer a replacement for the study of mathematics, science, and social science as discrete disciplines in the curriculum. Indeed, the vast majority of USMES teachers interviewed during 1973-74 accepted USMES as a supplement to regular class work, especially in mathematics and social studies. (More frequently, USMES was the replacement for other science instruction.) Teachers felt that mathematics skills in particular needed to be taught directly, and then they could be reinforced with the problem solving activities of USMES (Shann, 1975, p. 41). With the shifting claims about the role of USMES in the school's program came a new version of the perennial question about accountability for basic skill development: "Since time spent on USMES may detract from time spent on instruction in basic skills, will students in USMES classes

fall behind students in control classes in basic skill development?" Though the subjects in competition with USMES for instructional time became a clouded issue, interest in USMES students' basic skill development heightened.

Previous evaluations of USMES had shown no difference in the rates of basic skill development for USMES and control groups, as measured by the Reading Comprehension and Mathematics Computation subtests of the Stanford Achievement Test Series (Shann, 1975; Shapiro & Aiello, 1974). However, the usable data bases were small, and were not truly representative of the wider geographical and socioeconomic distributions of USMES users. In the interest of continuity, the SAT was selected for the 1974-75 program of data collection, and the previously administered reading and mathematics subtests were included in the expanded group of subtests administered to sample classes in 1974-75.

Sample Selection

Purposive sampling of new and experienced USMES teachers from 15 geographic areas was used to achieve a national sample of USMES classes representing a cross section of grade levels, socioeconomic levels, and unit challenges. Control classes came from non-USMES schools which were located in the same or neighboring communities as the USMES schools. These control classes were selected to match the USMES sample classes, one-for-one, on the bases of grade level, socioeconomic level, geographic area, and general features of the

schools' program. Field personnel were trained to serve as on-site evaluators for administration of the SAT subtests and for other kinds of data collection included in the evaluation program. The objective was to achieve complete, representative, and usable returns from a sample of 40 USMES classes and 40 control classes for a variety of student performance indices, observations, and teacher reports. The data base for analysis of the SAT's fell short of this goal.

Difficulties with Data Collection

One of the criteria for the original selection of the Stanford battery for the assessment of basic skill development was that this series was used more widely than other test batteries under consideration, and the technical features of the SAT were competitive. (Buros, 1972) Its selection would maximize the probability that at least some sample schools would be using it, and the requests for additional testing time for the USMES evaluation could be minimized.

Very few schools in the 1974-75 sample used the SAT as part of their regular school testing program, if they had one. Many principals and teachers, particularly those in the control sample, voiced objections to the time demands for the SAT administration. The problem was particularly acute for classes at the upper grade levels which faced strict schedules for the change of classes. Typically one class period provided more than enough time for the administration of one subtest, but it was insufficient for two. Either four separate class periods had to be used to administer four subtests, or schedule changes had to be made which affected many more students

than those in the evaluation sample. Despite these complaints, most of the sample schools did agree to cooperate in giving the SAT's for this USMES evaluation.

One notable exception was found in a school district where the earlier, 1964 edition of the SAT series was still being used, and the district, understandably, would not add to its already extensive testing schedule. Yet these scores from the 1964 versions of the SAT could not be used in the analysis of scores from the 1973 edition for the rest of the evaluation sample classes because equations for transformation of 1964 edition scores to their equivalents from the 1973 edition had not been available from the SAT publishers.

Additional problems diminished the data base for the basic skills assessment even further. A few teachers who were selected for the evaluation sample and who cooperated in other aspects of the testing program refused to allow administrations of the SAT's. In two geographical areas, which encompassed a total of 10 sample classes (USMES and control), state and local regulations on testing were severely limiting, and the requirements for securing parental permission could not be fulfilled by the evaluators working at a distance. Broken commitments by on-site evaluators or participating teachers accounted for pre-test data without any post-test results for a few additional classes. Finally, scores for two classes on two subtests had to be deleted from the analysis because the wrong subtests were administered at post-test time.

Six subtests of the SAT's were chosen for administration. As in prior years' evaluations, Reading Comprehension and Mathematics Computation

were administered to all sample classes in which the SAT's could be given. Four other SAT subtests whose content related more directly to the focus of the USMES program were introduced in the 1974-75 USMES evaluation plan. These were the subtests of Mathematics Application, Mathematics Concepts, Science, and Social Science. The skills and content areas these subtests are purported to sample are identified by the SAT publishers' test catalogue (Test Department, Harcourt, Brace, Jovanovich, Inc., 1973). To reduce the demands for testing time, one-half of the USMES classes and their controls received the Mathematics Concepts and Science subtests, while the other half of the sample received the Mathematics Application and Social Science subtests. The assignments were divided to ensure that all grade levels would be represented by all distributions of subtest scores. The numbers of classes for which usable pre- and post-test results were obtained are given for each subtest in Table 1. USMES and control classes were represented about evenly in the n's for each subtest.

Test Administration, Scoring, and Data Analysis

Pre-test and post-test administration of the SAT's was accomplished in the Fall (late September and early October) and in the Spring (during May) for both USMES and control classes. The tests were given by the trained field staff with the assistance of the classroom teachers who remained in their classrooms during the administration of these tests. Various levels of the test were employed, one per class, according to the guidelines provided by the SAT publishers. (Test Department, Harcourt, Brace, Jovanovich, 1973, p. 10). Some of the sample

classes, both USMES and control, from "non-graded" schools represented more than one grade. The level of the SAT administered in such classes was the level which was appropriate for the average of the given combination of grades. It was not feasible to administer more than one level of a subtest in a single classroom because testing times varied across levels, and oral directions differed as well. Students who received Primary I, II, or III level SAT booklets responded in the booklets, and their answers were transferred to Digitek answer sheets for optical scanning. The older students, who got Intermediate I, Intermediate II, or Advanced level tests, responded directly on the machine scorable answer sheets.

Scaled scores for each student on each of the subtests he took were obtained from scoring programs run at the Boston University Computing Center. The scaling procedure, which is described in the SAT norms booklets (Harcourt, Brace, Jovanovich, 1973, p. 13), is supposed to offer a system of inter-battery standard scores with comparability across levels for a test area. The scale values were derived by setting the median raw scores of grade 3 and grade 8 in their Fall standardization sample equal to 132 and 182 respectively.

As expected, there were in most classrooms some instances of loss of data from pre-test to post-test due to both student absenteeism and to the mobility of students. However, unlike former years' evaluations of USMES when the usable results were so limited that student rather than classroom had to be used as the unit of analysis, this evaluation could employ the classroom, which had been the sampling unit,

as the more appropriate unit of data analysis.

The data for each subtest were submitted to a repeated measures analysis of variance to investigate whether classes from either treatment group realized statistically significant gains in any of the six subtest areas from Fall to Spring. The two levels for the treatment factor included "USMES" classes, which worked on one or more USMES challenge at the teacher's discretion, and "control" classes, i.e., those non-USMES sample classes which received their regular school school program. Grade was also used as a factor in the design, because performance in the SAT subtest areas should be expected to increase over grade levels. Blocking on grades yielded three groups: grades 2-4; grades 5-6; and grades 7-8. These particular groupings were selected because they yielded reasonably large cell n's for an approximately proportional design (2:2:1), and none of the combination grade sample classes contained students whose grade placements overlapped the grade blocks, i.e., there were no classes of both fourth- and fifth-graders, nor of both sixth- and seventh-graders. The two-factor repeated measures analysis also enabled the investigation of several interaction effects.

Results

Results from the six repeated measures analyses of variance are summarized in Table 1. Complete ANOVA tables and tables of cell means

Insert Table 1 about here

and standard deviations from the analysis of each subtest are given in the final report to NSF on the evaluation of USMES (Shann, et al., 1975, 96-112). The actual results of these analyses conform closely to what one might expect. The highly significant increases ($p < .0001$) in all six subtest measures from the lower to high elementary grades for both treatment groups were consistent with expected growth patterns for these areas of achievement. Further, the significant growth patterns from pre-test to post-test administration which one would hope to find for both treatment groups within each grade level were found for all but one subtest, Mathematics Concepts. Those concerned that the USMES program no attempt to increase problem solving abilities at the expense of basic skill development or the acquisition of other knowledge and concepts were pleased to find no significant treatment differences favoring the non-USMES group. In fact, the observed treatment differences consistently favored USMES, especially at the upper grade levels, but these differences were not statistically significant.

Across the six analyses, 24 interaction effects were tested, but the vast majority (20) were not significant. The analysis for Reading Comprehension and for Mathematics Computation subtest scores each produced a significant grade by test administration interaction. In Reading Comprehension, the pre-post mean gains for the three grade blocks, 2-4, 5-6, and 7-8, were 12, 7, and 1 points, respectively. In Mathematics Computation, the pre-post mean gains by grade level were 12, 6 and 0.2. These patterns of reduced gains for the upper grade levels may be a function of a ceiling effect of these subtests and of lesser

attention devoted to these subjects over the years of a school's program. In addition, these two significant interaction effects mirror the growth curves for these basic skills.

The other two significant interaction effects reported in Table 1 were observed from the analysis of social science subtest scores. Although the USMES groups at all three blocks of grade levels showed gains from pre-test to post-test, their rate of growth declined for the seventh and eighth grades. The patterns of change for the control classes were much less consistent. The control group at the lowest block of grades 2-4, showed virtually no change in average performance from pre-to-post test time. The control classess in grades 5 and 6 averaged six points higher on the social science post-test than on the pre-test, and the control group at the seventh- and eighth-grade level witnessed a 7-point decline in group performance.

With the exception of the two interaction effects from the analysis of the social science subtest scores, the results from the repeated measures analyses presented no surprises; they conformed solidly with the results one should expect. Nonetheless, some problems with the data analysis should be noted.

If one accepts the use of analysis of covariance as a procedure for making statistical adjustments for pre-test differences between groups when randomized experiments are not feasible, and noted authorities express both positive (Cochran, 1957, p. 264; Ferguson, 1971, p. 288; Kerlinger, 1971, p. 373) and negative (Aiken, 1970; Cronbach & Furby, 1970) positions on the issue, then one might advocate its use here.

However, the covariance procedure assumes that the correct form of the regression equation has been fitted. The evaluators had anticipated that linear regressions would be appropriate, but continued investigations of the SAT data suggested that curvilinear equations may give a better fit for some SAT subtest data, more so for the control group than for the USMES group.

Another assumption required for the valid use of covariance is that error effects have a common variance, but the SAT data did exhibit departures from homogeneity of variance. Since this assumption is the natural extension of one required for analysis of variance, the problem of heterogeneity may affect the inferences one can draw even from the repeated measures analysis.

Investigations of the data suggest that questionable scaling procedures for SAT raw scores were responsible for the lack of homogeneity of variance. The evaluators were concerned that the formulae given by the SAT test developers really had not achieved comparability of translated, scaled scores across test levels for a given test area, but it was beyond the scope of the USMES evaluation to test this hypothesis directly. Drastic differences were observed in the sizes of the standard deviations for the cells in the two-factor pre-post design (see Shann *et. al.*, 1975, p. 114). The cell standard deviations for grade levels should have been consistently smaller than those presented in the norms booklet, because in this evaluation class means were used as the unit of analysis, and n's then were relatively small. Standard deviations for a given pre-test or post-test were comparable in size for USMES versus control groups at each grade level. However, the pre-test

versus post-test variances were very different for some grade levels on some of the post tests, and there were most notable inconsistencies in the sizes of the standard deviations across grade levels. The very large standard deviations were found principally in the block of grades, 2-4. This observation drew attention to the fact that the data for this block of grades did differ in one important respect from the data for grades 5-6 and for grades 7-8. The levels of the SAT which were prescribed for pre-test and then post-test administration to classes in grades 2-4 encompassed four levels of the SAT: Primary I, II, and III, through Intermediate I. For grades 5 and 6, only two levels were needed for pre-test and post-test administration: Intermediate I and II. In the block for grades 7 and 8, only the one Advanced level was required.

Scaled scores for the SAT were developed in such a way that this use of various levels of the sequential battery should not have made a difference. The test developers claim that their scaling procedure resulted in a system of inter-battery standard scores which should permit the translation of raw scores at each level to standard scores with comparability across levels for a test area. Thus, when the appropriate conversion tables were used, the choice of test level, within the narrow limits offered for a grade level, should not have made a difference in the scaled score which an individual obtains. The observation of extremely varied standard deviations among the cells of this two-factor, repeated measures design, especially among grade blocks and between test administrations, may threaten the validity of this claim.

Discussion

The seriousness of the problem of heterogeneity of variances in the SAT data for the USMES evaluation was diminished greatly by the reasonableness and the consistency of the overall results of the analysis of basic skill development over previous evaluations and within this evaluation. Graphs of cell means given in the final evaluation report (see Shann, et al., 1975, pp. 118-123) illustrate very pointedly the magnitude of the expected grade level differences and the proximity of the means for the treatment groups at each grade level on both pre-tests and post-tests. Sophisticated analyses were not needed to ascertain that the national sample of USMES classes performed at least as well as their control matches on the six selected subtests of the Stanford Achievement Test battery.

Other vehicles of data collection for the USMES evaluation, notably teacher interviews and questionnaires, yielded information salient to the finding of no treatment differences in performance on the SAT's. The teacher data revealed that USMES students had not been deprived of instruction in the basic skills. In some cases, they may have been getting more than non-USMES students. Without exception, for all 40 sample USMES teachers interviewed, the teachers continued to teach mathematics as a separate content area; their students were getting their "ordinary" mathematics instruction plus "USMES math." As for science, in the majority of sample classes, USMES was the science program, for a portion of the school year, or for the entire school year, and one wonders if there would have been any science in these classes if USMES

were not presented. Thus, in the mathematics and science skills area, USMES should not have been interfering with skill and concept development, but rather, adding to it.

In other skill areas, i.e., reading, language arts, and social science, reports from teachers support the conclusion that USMES was not taking much, if any, time from these areas. Especially for classes which were not self-contained, schedules and amounts of time were mandated for instruction in these areas. In light of the patterns of USMES usage in most schools, the issue of basic skill development should be of diminished importance to current and prospective USMES users. More pointedly, the issue is no longer an issue.

Standardized achievement test batteries continue to reflect emphasis on basic skills and factual knowledge, though there is a movement toward measuring more application and interpretation. If the test content is in harmony with the primary objectives of a new curriculum, then it may be worthwhile to invest the effort and resources necessary for using a good standardized test properly. The primary objective of the USMES program has been the enhancement of elementary school students' skills in complex problem solving. Yet, despite documentation that USMES use was not interfering with basic skill instruction, it was unlikely that fears about the impairment of basic skills of USMES students could have been allayed so readily without the SAT results reported here.

In addition to the costs, the time requirements, and the difficulties encountered in data collection with the SAT's, the experiences of the basic skills assessment in the USMES evaluation offer another point for educational researchers, evaluators, and testing directors

to consider: are the scaled scores comparable across test levels in a battery which purports to be sequential? If such a claim is not valid, then comparisons across grade levels are suspect, longitudinal studies with the measures are futile, and data supposedly reflecting an individual's progress in basic skills are meaningless. One way to test the comparability of scaled scores across test levels would involve administering, close in time, three adjacent levels of the same subtest to a group of subjects for whom the middle level of difficulty is considered most appropriate. Several groups of subjects should be chosen so that their assigned test levels would overlap across the range of levels offered by the test series.

A truly rigorous assessment of accountability within a cost-analysis framework could not be undertaken for USMES. If basic skill development were the primary objective, then the program might be deemed a failure, not to have significantly increased the rates of basic skill development for USMES over non-USMES students. Very substantial human and financial resources have been expended to support the development and trial implementation of the USMES program. However, competencies in real, complex problem solving were the primary concern of USMES, not basic skills. The final evaluation report (Shann, et al., 1975) documented the decided perception of the USMES teachers that the USMES program does teach problem solving skills to its students, but the more objective available instruments to measure problem solving skills were still too unsophisticated to give an accurate reading of this same question. New instrument development for complex problem solving became a dual thrust of the USMES

evaluation team's efforts, but the new techniques were not ready for use in the evaluation, only for methodological research on the instruments themselves (see Shann, 1976). A valid accounting of the program's worth is dependent on a valid measure of complex problem solving.

References

- Bell, Terrel. The Means and Ends of Accountability. Proceedings of the Conferences on Educational Accountability. Princeton, N.J.: Educational Testing Service, 1971.
- Buros, O. K. The Seventh Mental Measurements Yearbook. Highland Park, N.J.: Gryphon Press, 1972.
- Catalogue of tests and related services. Test Department, Harcourt, Brace, Jovanovich, Inc. 1973.
- Cunningham, Lavern L. Our Accountability Problems. In L. H. Browder, Jr. (Ed.). Emerging Patterns of Administrative Accountability. Berkeley, Calif.: McCutchan Publishing Corp., 1971.
- Dyer, Henry S. Toward Objective Criteria of Professional Accountability in the Schools of New York City. Phi Delta Kappan, 1970, Vol. 52, No. 4, 206-211.
- Educational Development Center. Unified science and mathematics for elementary schools. Newton, Mass.: September, 1972.
- Gallup, George. The Third Annual Survey of the Public's Attitude toward the Public Schools. Phi Delta Kappan, 1971, Vol. 53, No. 1, 33-48.
- Gooler, Dennis D. Some Uneasy Inquiries into Accountability. in L. M. Lessinger and R. W. Tyler (Eds.). Accountability in Education. Worthington, Ohio: Charles A. Jones, 1971.
- Locke, Robert W. Accountability Yes, Performance Contracting Maybe. Proceedings of the Conferences on Educational Accountability. Princeton, N.J.: Educational Testing Service, 1971.

- Mehrens, William A., & Lehmann, Irvin J. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart & Winston, 1973.
- Shann, Mary H. An evaluation of Unified Science and Mathematics for Elementary Schools (USMES) during the 1973-74 school year. Interim report to the National Science Foundation on grant no. SED 74-00542 A01. Boston University, August, 1975.
- Shann, M. H. Measuring problem solving skills and processes of elementary school children. Report on new instrument development to the National Science Foundation. Boston University, June 1976.
- Shann, M. H., et al. Student effects of an interdisciplinary curriculum for real problem solving: The 1974-75 USMES evaluation. Final report to the National Science Foundation on grant no. SED 74-00542 A01. Boston University, December, 1975.
- Shapiro, B. J. & Aiello, T. Report on basic skill development. Report to the National Science Foundation. Boston University, June, 1974.
- Stanford Achievement Test, Form A, Norms Booklet. Harcourt, Brace, Jovanovich, Inc., 1973.

Table 1
Summary of Significant Results of Repeated Measures
Analyses of Variance on Six SAT Subtests

Subtest	n ^a	Treatment ^b	Grade ^c	Pre-Post ^d	Interactions
Reading Com- prehension	50	N.S.	$p < .0001$	$p < .0001$	G X A ^e , $p < .05$
Mathematics Computation	58	N.S.	$p < .0001$	$p < .0001$	G X A ^e , $p < .0001$
Mathematics Application	22	N.S.	$p < .0001$	$p < .01$	none
Mathematics Concepts	31	N.S.	$p < .0001$	N.S.	none
Science	29	N.S.	$p < .0001$	$p < .01$	none
Social Science	25	N.S.	$p < .0001$	$p < .01$	G X A ^e , $p < .05$ T X A ^f , $p < .05$

^an = number of classes; class means were used as the unit of analysis

^bUSMES versus non-USMES

^cGrades were grouped into three blocks: 2-4; 5-6; and 7-8.

^dThese were Fall and Spring administrations of the SAT subtests.

^eInteraction of grade with test administration

^fInteraction of treatment with test administration